D(PROG) / DT : SENSE OR NONSENSE?

Thomas M. Hamill

University of Colorado and NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado

7 June 2002

submitted as a NOTE to *Weather and Forecasting*

DRAFT

Corresponding author address: Dr. Thomas M. Hamill, NOAA-CIRES CDC, R/CDC 1, 325 Broadway, Boulder, CO 80305-3328. e-mail: hamill@cdc.noaa.gov

ABSTRACT

Forecasters often use the term “$d$prog/dt” to discuss changes in numerical forecasts verifying at the same time. For example, if yesterday’s 48-h temperature forecast was warm and today’s 24-h forecast is cooler, $d$prog/dt is negative. Some forecasters may also think of $d$prog/dt as a general rule of thumb for how to improve upon the latest model forecast. Given a set of lagged-average forecasts from the same model all verifying at the same time, this rule of thumb suggests that if the forecasts show a trend, this trend is more likely than not to continue and thus provide useful information for correcting the most recent forecast. Forecasters may also note the amount of continuity of forecasts as a judge of the magnitude of the error in the most recent forecast.

To examine this rule of thumb, a 23-year record of forecasts are generated from a T62 version of the medium-range forecast model used at the National Centers for Environmental Prediction. Forecasts are initialized from reanalysis data, and January-February-March forecasts are examined for selected locations. The validity of “$d$prog/dt” is assessed using 850 hPa temperature forecasts. Extrapolation of forecast trends are shown to have little forecast value. Also, there is only a small amount of information on forecast accuracy from the amount of discrepancy between short-term lagged-average forecasts.

1. INTRODUCTION

Numerical weather forecast models grow increasingly sophisticated with each passing year. Unfortunately, the quest for a model free of systematic error remains elusive. Weather forecasters often develop rules of thumb to adjust the guidance produced by numerical weather prediction models. Nonetheless, human forecasters are fallible; their rules may appear to be appropriate from a relatively small sample of recent forecasts. However, their ability to validate their algorithms with a longer time series of forecasts is complicated by the frequent coding changes to the weather forecast models; as soon as they feel they are starting to understand the bias of one version of the model, that model is changed and the learning process must start over.

Given these model changes, a rule of thumb that can be applied regardless of the specific forecast model would be especially valuable. One potentially fruitful avenue for improving upon the latest numerical guidance is to consider multiple forecasts from the same model valid at the same time. Such *lagged-average forecasts* ("LAFs"; Hoffman and Kalnay 1983, Dalcher et al. 1988) have previously been shown to be useful for improving the skill of medium-range forecasts. For shorter-range forecasts, an evaluation of trends in lagged-average forecasts is often referred to as "$d$prog/dt." Thus, one may see in forecast discussions that "temperature $d$prog/dt is negative," meaning that more recent numerical forecasts are colder than older ones. Some forecasters may also view $d$prog/dt as a handy rule of thumb: if forecasts are trending colder, does that not suggest that the most likely actual state is yet somewhat colder than the most recent forecast? Forecasters may also note the amount of continuity of these lagged-average forecasts as a judge of the likely error in the most recent forecast. Lagged-average forecasts that have been consistent are judged to be more accurate than ones that substantially differ from each other.

Is "$d$prog/dt" as a rule of thumb justified by the data? Answering this question in a statistically rigorous manner has been complicated by the unavailability of a long time series of forecasts produced by the same forecast model. At the Climate Diagnostics Center (CDC), we have initiated a "re-forecasting" project to study whether significant improvements to forecast skill are possible if a very long time series of forecasts are available from a frozen model, where the model algorithms are unchanged throughout the experiment. Using this large training data set, systematic model errors can be detected, and current forecasts using the same frozen model can be adjusted for these errors. We have thus far generated 23 years of medium-range weather forecasts from a T-62 resolution version of the National Center for Environmental Prediction's (NCEP's) medium-range forecast (MRF) model (Kanamitsu 1989, Kanamitsu et al. 1991, Caplan et al. 1997, Wu et al. 1997). A single control forecast has been run forward for 2 weeks once every day from 0000 UTC initial conditions using the NCEP-NCAR reanalyses (Kalnay et al. 1996) from 1979 to 2001. Over the next year, we expect to complete a 15-member ensemble of forecasts over the 23 years. This data set will primarily be used for demonstrating how medium-range forecasts can be improved. However, it can also be useful for a range of ancillary applications, such as the validation of rules of thumb like $d$prog/dt.

Ideally, the most appropriate data to test would be the ones forecasters are now using. Hence, if forecasters are applying $d$prog/dt to 12, 24, and 36-h forecasts from the state-of-the-art forecast model at NCEP, a model such as the Eta model (Black 1994, Rogers et al. 1995, 1996, Mesinger 1996) would be the best to test. However, the Eta model is frequently modified and improved at NCEP, so a long history of forecasts using the same model configuration is not available. Still, $d$prog/dt is generally thought of as being a tool which may be used with different numerical models, so an assessment of its validity with a simpler model should provide valuable guidance about whether it should be used with current operational forecast models.

The rest of the note consists of a brief examination of the skill of this forecast data set, an examination of how much improvement can be obtained through lagged-average regression approaches, and an examination of the validity of the $d$prog/dt rules of thumb.

2. RESULTS

Our data set will consist of 1, 2, and 3-day forecasts of 850 hPa temperature from January-March 1979-2001. Sea-level pressure forecasts were also examined but will not be shown here;

the results were both qualitatively and qualitatively similar. Associated NCEP-NCAR reanalysis will be used as verification data. For simplicity, we will examine the validity of $d$prog/dt at a limited set of locations in the U.S.. These locations are the grid points nearest to Seattle, WA, Los Angeles, CA, Denver, CO, Minneapolis, MN, San Antonio, TX, Columbus, OH, Tampa, FL, Cape Hatteras, NC, and Portland, ME. To minimize the direct effect of forecast bias and the annual cycle upon the analysis, a 31-day running mean climatology of the analysis state and the mean forecast state is computed for each of these locations using the full 23-year data set. These running means are subtracted from the analysis and forecasts prior to the subsequent examination.

a. Validity of extrapolating forecast trends

First consider the overall error statistics of these forecasts. Table 1 provides the root mean-square (RMS) error characteristics of the forecasts at the nine locations as a function of lead time.

As a baseline for evaluating the value of forecast trends, we perform a simple univariate regression to predict the 850 hPa temperature provided with just the 24-h forecast temperature. For this regression, a cross-validation approach is used (Wilks 1995). The regression constants are separately calculated for each of the 23 years, using the remaining 22 years as training data. Denote T_{pred} as the predicted 850 hPa temperature (deviation from observed climatology) and T_{24} the 24-h forecast (deviation from forecast climatology). The regression equation is of the form

$$T_{pred} = \beta_0 + \beta_1 \times T_{24} \tag{1}$$

The RMS error of this univariate regression is also displayed in column 5 of Table 1. The errors are consistently slightly lower than those from the 24-h forecast itself.

If there is value in the trend in lagged-average forecasts, inclusion of these trends ought to significantly improve the accuracy of these forecasts. Accordingly, denote $(T_{48} - T_{24})$ the trend between 48 h and 24 h lagged average forecasts valid at the same time, and similarly for $(T_{72} -$

T_{48}). We then perform a cross-validated multivariate linear regression of the form

$$T_{pred} = \beta_0 + \beta_1 \times T_{24} + \beta_2 \times (T_{48} - T_{24}) + \beta_3 \times (T_{72} - T_{48}) \tag{2}$$

The RMS errors of this multivariate regression are also displayed in the last column of Table 1. The inclusion of additional information on forecast trends makes only a very small improvement to the skill of the forecasts; on average, only 0.04 K. If one examines the distribution of regression coefficients produced via the cross-validation (not shown), the distribution of β_2 and β_3 typically overlap zero, indicating little confidence that the optimal values for these coefficients are significantly different from zero.

Examining a scatterplot of 48-24 h forecast trends and their relationship to the difference between the 24 h forecast and the analyzed state, the reason for the limited value of extrapolating trends is more apparent. Figure 1 provides this scatterplot; the difference in temperatures between 48 and 24-h forecasts valid at the same time is plotted along the x-axis, the difference between 24-h forecasts and the verification along the y-axis. There is little relationship between the forecast trend and the 24-h forecast error, as noted by the correlation coefficients near zero (plotted in the upper-left corner of each panel). The correlations were generally smaller yet if the trend was evaluated between 72 h and 24 h.

b. Estimating forecast skill from consistency

Is the consistency of forecasts useful for inferring something about the accuracy of the most recent forecast? Figure 2 provides a scatterplot of the absolute difference between 48 and 24 h forecasts (x axis) and the mean absolute error (MAE) of the 24 h forecasts (y axis). Ideally, the larger the discrepancy between 48 and 24 h forecasts, the larger the typical MAE should be. Let F denote the absolute difference between 48 and 24 h forecasts. On Fig 2, we also plot $\overline{MAE(F \leq 1)}$, $\overline{MAE(1 < F \leq 3)}$, and $\overline{MAE(3 < F)}$, where the overbar denotes the average over all forecasts. Note that there is typically very little difference between the average MAEs of forecasts with large discrepancies and small discrepancies.

3. CONCLUSIONS

$d$prog/dt has been shown to have little validity as a forecast rule of thumb. Temperature trends with this model should not be extrapolated, and there is only a little value in the amount of discrepancy in lagged average forecasts for predicting the magnitude of forecast error.

Is this apparent lack of improvement a consequence of using this particular model? While rules of thumb are often model-dependent, this particular rule seems to be applied regardless of the model. Following this same reasoning, $d$prog/dt should be considered useless for other forecast models until and unless a careful validation demonstrates its utility.

There are demonstrably valuable techniques for estimating forecast uncertainty and improving the skill from a single deterministic forecast. One such technique is commonly referred to as *ensemble forecasting* (Toth and Kalnay 1993, 1997; Molteni et al. 1996; Houtekamer et al. 1996). There is a smaller body of literature on the usefulness of ensembles for shorter-range forecasts. See Brooks et al. (1992) for a motivation for short-range ensemble forecasting and Hamill et al. (2000), for a literature review. There are many challenging problems that need to be addressed to improve these forecasts, but such data sets should be more useful for evaluating the uncertainty of shorter-range forecasts. Readers who may have used $d$prog/dt but are looking for a more theoretically justifiable alternative are encouraged to consider the information from these ensembles.

4. ACKNOWLEDGMENTS

Richard Grumm (National Weather Service, State College, PA) is gratefully acknowledged for his consultation during the drafting of this manuscript.

REFERENCES

Black, T. L., 1994: The new NMC mesoscale eta model. description and forecast examples. *Wea. Forecasting*, **9**, 265-284.

Brooks, H. E., C. A. Doswell, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting*, **7**, 120-132.

Caplan, P., J. Derber, W. Gemmill, S.-Y. Hong, H.-L. Pan, and D. Parrish, 1997: Changes to the 1995 NCEP operational medium-range forecast model analysis-forecast system. *Wea. Forecasting*, **12**, 581-594.

Dalcher, A., E. Kalnay, and R. N. Hoffman, 1988: Medium-range lagged average forecasts. *Mon. Wea. Rev.*, **116**, 402-416.

Hamill, T. M., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range. Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653-2664.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte-Carlo forecasting. *Tellus*, **35a**, 100-118.

Houtekamer, P. L. , L. Lefaivre, and J. Derome, 1996: The RPN ensemble prediction system. *Proc. ECMWF Seminar on Predictability, Vol II*, Reading, United Kingdom, 121-146. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom].

Kalnay, E., and co-authors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437-472.

Kanamitsu, M., 1989: Description of the NMC global data assimilation and forecast system. *Wea. Forecasting*, **4**, 334-342.

———, and Coauthors, 1991: Recent changes implemented into the global forecast system at NMC. *Wea. Forecasting*, **6**, 425-435.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bull. Amer. Meteor. Soc.*, **77**, 2637-2650.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73-119.

Rogers, E., D. G. Deaven, G. S. Dimego, 1995: The regional analysis system for the operational "early" Eta model: original 80-km configuration and recent changes. *Wea. Forecasting*, **10**, 810-810.

——— , T. L. Black, D. G. Deaven, G. J. DiMego, Q. Zhao, M. Baldwin, N. W. Junker, Y. Lin, 1996: Changes to the operational "early" Eta analysis/forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting*, **11**, 391-416.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317-2330.

——— , and ——— , 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **12**, 3297-3319.

Wilks, D. S., 1995: *Statistical methods in the atmospheric sciences: an introduction.* Academic Press. 467 pp.

Wu, W. M. Iridell, S. Saha, and P. Caplan, 1997: Changes to the 1997 NCEP operational MRF model analysi/forecast system. *NCEP Technical Procedures Bulletin 443*, available from http://www.nws.noaa.gov/om/tpb/indexb.htm, or the National Weather Service, Office of Meteorology, Programs and Plans Division, Silver Spring, MD, 20910.

FIGURE CAPTIONS

Figure 1. Validity of $d$prog/dt for various locations. Plotted on x-axis in each panel are the 23 years $\times$ 90 days of 850 hPa temperature differences between 48-h and 24-h forecasts valid at the same time. On the y-axis are differences between the 24-h forecast and the verification. Correlation coefficient plotted in upper-left corner.

Figure 2. Mean absolute error (MAE) of 24-h 850 hPa temperature forecasts (y-axis) and their relationship to the absolute difference between 48-h and 24-h temperature forecasts (x-axis, forecasts valid at the same time). Horizontal lines denote the average MAE for cases when the absolute difference was less than 1 C, between 1 and 3 C, and greater than 3 C.

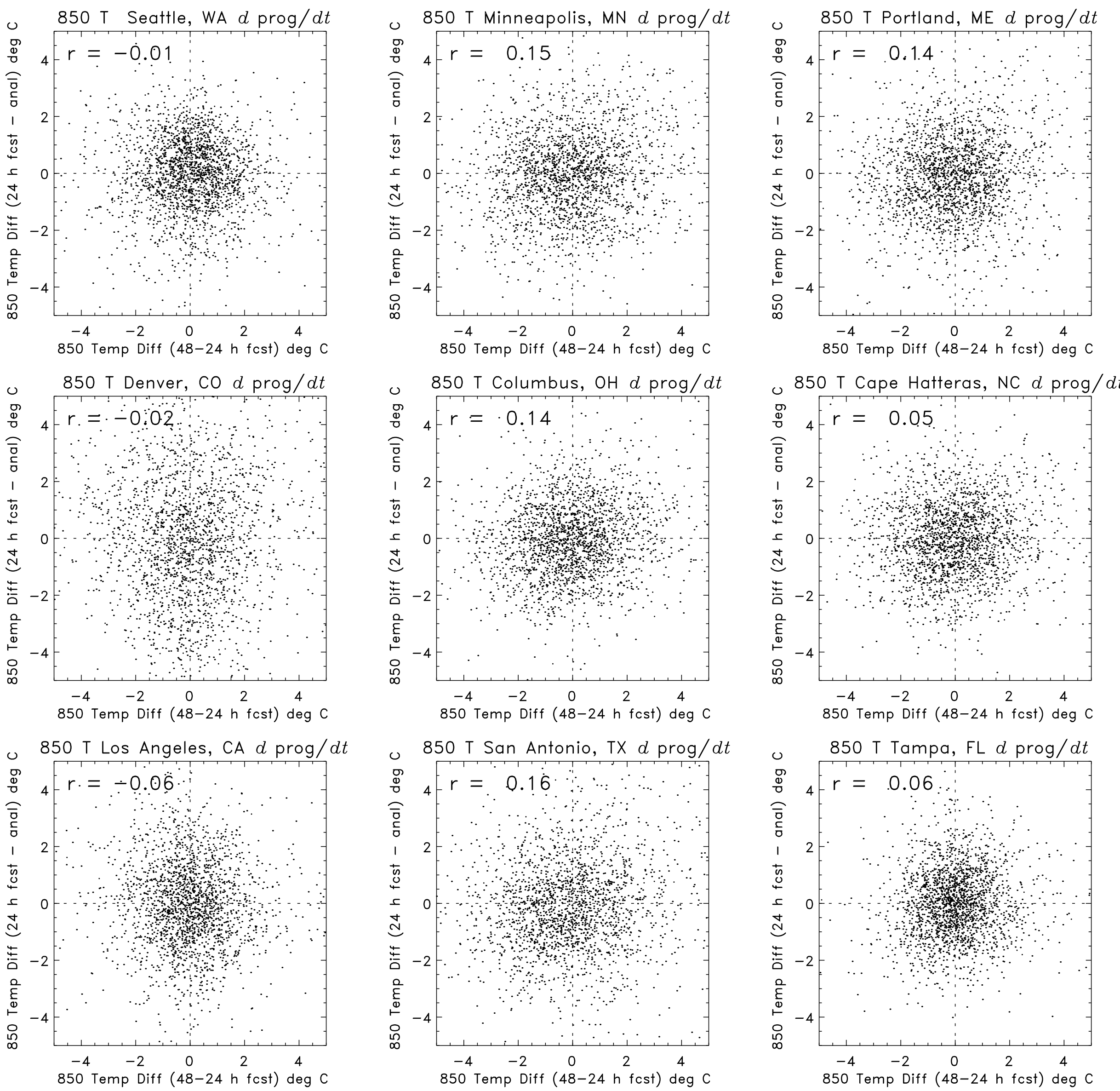

Figure 1. Validity of $d$prog/dt for various locations. Plotted on x-axis in each panel are the 23 years × 90 days of 850 hPa temperature differences between 48-h and 24-h forecasts valid at the same time. On the y-axis are differences between the 24-h forecast and the verification. Correlation coefficient plotted in upper-left corner.

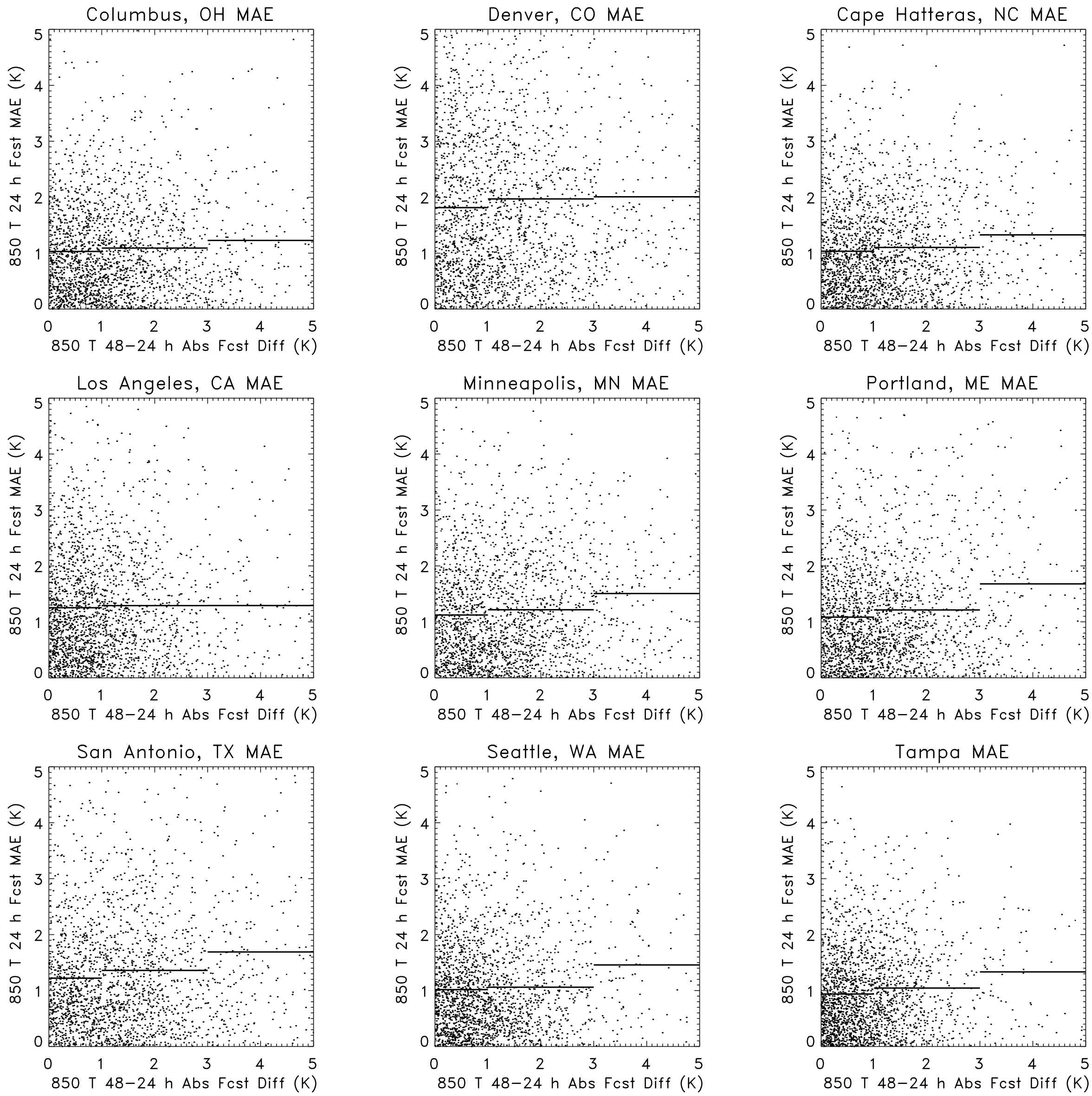

Figure 2. Mean absolute error (MAE) of 24-h 850 hPa temperature forecasts (y-axis) and their relationship to the absolute difference between 48-h and 24-h temperature forecasts (x-axis, forecasts valid at the same time). Horizontal lines denote the average MAE for cases when the absolute difference was less than 1 C, between 1 and 3 C, and greater than 3 C.

Table 1. 850 hPa RMS temperature errors (degrees C) for selected locations. First column denoted the location; columns 2, 3, and 4 denote the RMS errors of 24, 48, and 72 h forecasts, respectively. Column 5 indicated the RMS error of a linear regression forecast with one predictor (24-h 850 hPa temperature), and column 6 the error of a multivariate linear regression forecast with 3 predictors (24-h 850 hPa temperature, 48-24 h difference, and 72-48 h difference).

	24-h RMSE	*48-h RMSE*	*72-h RMSE*	*Regr1 RMSE*	*Regr3 RMSE*
Seattle, WA	1.36	2.01	2.60	1.31	1.26
Denver, CO	2.43	3.17	4.01	2.22	2.16
Los Angeles, CA	1.64	2.19	2.71	1.33	1.29
Minneapolis, MN	1.57	2.68	3.79	1.53	1.49
Columbus, OH	1.39	2.43	3.53	1.36	1.34
San Antonio, TX	1.80	2.90	3.82	1.75	1.69
Portland, ME	1.59	2.51	3.67	1.56	1.51
Cape Hatteras, NC	1.39	2.25	3.56	1.35	1.34
Tampa, FL	1.27	1.87	2.62	1.21	1.18